Metadata evaluation, consistency, compliance and improvement **[1]**

Submitted by edward.m.armstrong on Mon, 2014-09-22 16:15 Tuesday, January 6, 2015 - 15:30 to 17:00
Event: Winter Meeting 2015 [2]
Session Type: Breakout [3]
Expertise Level: Beginner [4]
Collaboration Area: Documentation [5]
Information Quality [6]
Preservation and Stewardship [7]
Products and Services [8]
Abstract/Agenda:
This session will focus on tools and approaches for the evaluation and improvement of metadata from the perspective of error, consistency and quality. Both concrete examples (existing tools and services) and abstract ideas for new services needed are encouraged.

Notes:
Design of Community Resource Inventories as a Component of scalable Earth Science Infrastructure – Ilya Zaslavsky

· Catalogues and domain inventories

· Started by asking data centers what they would value most

·
http://www.geoportal.rlp.de/mapbender/php/mod_dataISOMetadata.php?outputFormat=iso19139&id=07d93d442d634bd4a991541b188daa8f&validate=true [9]

· Not just publish metadata but try to enhance – populate by parsing abstract and/or text in documents – extract key words and spatial extent

o CINERGI metadata harvesting

o Will keep provenance of what was done to the record

· Get lots of data where no one thought about metadata – can't get back to the originator to improve

o Have various harvesters – then federate schemes … auto enhance

· Enhancements – use API enhancer – map GCMD key words against record – end up with metadata documents that have a lot more key words

· Have metadata before and after that is compared

o Have provenance and extracted facets based on key words and JSON rdf file

· Have more than 1 interface where we can browse by different vocabularies

· Will try to work with different geosciences communities – can create sandbox to create resources

· Inventories is key in defining geoscience CI – how different communities can come together

· If data facilities – interested in completeness and quality of metadata – then use this to test

Metadata Metrics: History and Lessons Learned – Anna Milan

· NOAA metadata evaluation tool - EMMA

o Metrics = completeness, schema validation broken URLs count of components

· Started recording metrics when converted to ISO – top of metrics is 55,000 – red is validation errors

o Seen increase in records able to provide feedback to authors and senior personnel

· Basic metrics = # records – provides info to managers

· Histories – show how collection has improved over time

· Can have good scores (over 25 out 42) – one person manual cleaned up metadata

· Consistency Checker – checks how often a string is represented in fields – recognize inconsistency in metadata

· Completeness Rubric – looks at individual record – based on spiral – measures how complete or incomplete a record is

· Evaluation is good – but still need human insight to improve it

· Questions – how asses the quality? How can we measure count of datasets without metadata (only count things that exist), I want metrics for MY datatype, know all access points

· Observations

o Self-conscious about poor completion results – look like black mark on their work

o Mangers like quick overview of statistics

o Completeness measurements – means authors look at how to measure attributes that they normally overlook

o People put non-sense to get A++

o One size does not fit all

o Valuable visualization tool

o Still need human intervention to ensure meaningful content

· Do – engage community, put content in rubric assessment, simplify assessment results – don’t ignore the variety of data types and their uniqueness equate completeness with quality

· Q – is anyone tracking how use of datasets correlates to the metadata – No – it is a common concern

QA Rules – Tyler Stevens

· Metadata QA Process

o GCMD – CMR (common metadata repository) – schema evaluation, human intervention review (things beyond QA), changes made to metadata to improve – done by provider, notification of metadata changes, then published

· QA rules

o Accuracy, completeness, consistency, conciseness, readable/understandable

· Checks include: controlled vocabulary validity, field lengths uniqueness, required fields populated

· Rules are driven by: - use UMM-C (Unified metadata model for collections)

o Formats, models, requirements, experiences

· Rules include: link, character, date, numeric (field type), controlled vocabulary, miscellaneous (existing checks)

· QA rules can assist in assessing/improving metadata, can help automate some of the process, and engage the community

· Q – what are you using schema-tron to validation – no – first is a schema validation and then based on the rules

· Q – why not use schematron – doesn't know – will take that back for the testing process

· Q – is it available for people to use/test – no – has to go through NASA process

Metadata Compliance and Consistency Validation - Ed Armstrong, Oliver Change, Dave Foster (JPL)

· Tool developed by Oliver works at the granule level (not the collection level)

· Details are not just important for human data user – but the autonomous software systems – needed to be able to use them in visualization packages

· Popular metadata stands are CF and ACDD

· Public metadata checkers – CF Project compliance checker (puma.nerc.ac.uk/cgi-bin/cf-chcker.pl)

o UDDC (THREDDS) compliance checker – thredds.jpl.gov/thredds/uddc/ncml/aggregation/

o GHRSST compliance checker – command line tool – (PO.DAAC)

· IOOS Compliance checker – difficult – target dependencies, tied to a terminal output

· Took open source software to make a thin wrapper around the tool – html based

· Rewrote ACDD and GDS2 checker tool, left most of the CF checker as it was

· Upload a local granule OR use OPeNAP url – takes a few seconds to 2 min to check – results page is grouped by hierarchies (100s of tests are being performed), it is colored for pass/some/fail (green, yellow, red). – there is also a % score for each granule

· There is an api – can execute with a curl command with NetCDF url – get JSON output

· Not publically assessable right now – trying to get up on po.daac “labs”

Session Leads:

Name: Ed Armstrong [10]
Organization(s): NASA [11] ,Jet Propulsion Lab [12]
Email: edward.m.armstrong@jpl.nasa.gov [13]

Presenters:

Name: Ed Armstrong [10]
Organization(s): NASA [11] ,Jet Propulsion Lab [12]
Email: edward.m.armstrong@jpl.nasa.gov [13]

Name: Anna Milan [14]
Organization(s): NOAA [15] ,NGDC [16]

Name: Ilya Zaslavsky [17]
Organization(s): San Diego Super Computing Center [18]

Name: Tyler Stevens [19]
Organization(s): NASA Global Change Master Directory [20]

Notes takers:

Name: Kelly Monteleone [21]
Organization(s): TERA [22] ,a CH2M Hill Company [23]
Email: kalcan83@hotmail.com [24]

Participants:
Byron, Dan, Fan Feng, Jacqueline (JAci) Mize, Katie Baynes, Robert Wolfe, Peng, Aaron Sweeney, Ted Haberman, Alexsander Jelenek, Anna Milan, Ed Armstrong, KAi Liu, Ziheng Sun, Lisa Zolly, John Scialdone, Tyler Stevens, Jennifer Wei, Lisa Booker, Ajay Krishnan, Nate James, Sean gordon, Ellen Johnson, Barbara Brooks, Chris Torbert, Thomas Huang, Andrew Mitchell, Joe Lee, Lindsay Power, STeve Kempler, Pam Mlynczak (possibly more - only found 1 of 3 sign in sheets).

Teaser: Focus on tools and approaches for the evaluation and improvement of metadata from the perspective of error, consistency and quality.
Accepted:

Source URL: http://commons.esipfed.org/node/2684

Links:
[1] http://commons.esipfed.org/node/2684
[2] http://commons.esipfed.org/2015WinterMeeting
[3] http://commons.esipfed.org/session-type/breakout
[4] http://commons.esipfed.org/taxonomy/term/260
[5] http://commons.esipfed.org/collaboration-area/documentation

[6] http://commons.esipfed.org/collaboration-area/information-quality
[7] http://commons.esipfed.org/collaboration-area/preservation-and-stewardship
[8] http://commons.esipfed.org/collaboration-area/products-and-services
[9] http://www.geoportal.rlp.de/mapbender/php/mod_dataISOMetadata.php?outputFormat=iso19139&id=07d93d442d634bd4a991541b188daa8f&validate=true
[10] http://commons.esipfed.org/node/508
[11] http://commons.esipfed.org/taxonomy/term/228
[12] http://commons.esipfed.org/taxonomy/term/197
[13] mailto:edward.m.armstrong@jpl.nasa.gov
[14] http://commons.esipfed.org/node/2013
[15] http://commons.esipfed.org/taxonomy/term/242
[16] http://commons.esipfed.org/taxonomy/term/267
[17] http://commons.esipfed.org/node/533
[18] http://commons.esipfed.org/taxonomy/term/317
[19] http://commons.esipfed.org/node/285
[20] http://commons.esipfed.org/taxonomy/term/218
[21] http://commons.esipfed.org/node/7744
[22] http://commons.esipfed.org/taxonomy/term/1754
[23] http://commons.esipfed.org/taxonomy/term/1755
[24] mailto:kalcan83@hotmail.com